0.1. *Болдаков В.С.* Синтез речи с использованием векторных представлений эмоций

Многие люди предпочитают использовать голосовой интерфейс в своей повседневной жизни, например управление бытовой техникой при помощи умной колонки, или создания таймера через голосового ассистента смартфона. Такой интерфейс предполагает необходимость взаимодействия с пользователем посредством голоса. Чтобы не ограничивать функциональность программного обеспечения, используя предзаписанные человеком фразы, необходим синтез речи.

В последние годы появилось большое число исследований нейросетевых методов для качественного синтезирования речи. Например, авторегрессионная архитектура Tacotron 2 [1] или архитектура на базе трансформера FastSpeech 2 [2]. В указанных работах синтезируется естественная речь, но нет возможности контролировать эмоции, с которыми произносится текст.

Существуют решения, позволяющие при синтезе задавать конкретную эмоцию из ограниченного дискретного распределения. В данной работе описывается новый метод синтеза речи с широким спектром эмоций из непрерывного распределения векторных представлений $\mathbb{R}^{2304}$, полученных из текста с помощью модели, описанной в [3] и их последующего линейного преобразования и нормирования:

$$\boldsymbol{y} = \frac{\boldsymbol{W}\boldsymbol{x} - \mathbf{E}[\boldsymbol{W}\boldsymbol{x}]}{\sqrt{\mathbf{D}[\boldsymbol{W}\boldsymbol{x}] + \epsilon}}\gamma + \beta,$$

где $\boldsymbol{x}$ — полученное векторное представление текста, $\boldsymbol{W}$ — обучаемая матрица линейного преобразования, β, γ — обучаемые параметры.

В результате данной работы получено решение на базе архитектур Tacotron 2 и FastSpeech 2, позволяющее синтезировать эмоциональную речь. Данный способ получения эмоционального синтеза может быть применен к большинству существующих нейросетевых архитектур синтеза речи.

Научный руководитель — к.т.н. Ракитский А. А.

Список литературы

[1] Shen J., Pang R., Weiss R., et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions // Proc. Intern. Conf. «2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)». Calgary, AB, Canada: IEEE, 2018. P. 4779–4783.

[2] Ren Y., Hu C., Tan X., et al. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. [Электронный ресурс]. URL: https://arxiv.org/abs/2006.04558 (дата обращения 09.09.2021).

[3] Felbo B., Mislove A., Sogaard A., et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm // Proc. Intern. Conf. «Conference on Empirical Methods in Natural Language Processing». Copenhagen, Denmark: Association for Computational Linguistics, 2017. P. 1615–1625.